Causality and audio-visual integration

Michael Schutz and Michael Kubovy
University of Virginia

Schutz and Lipscomb (2007) reported an audio-visual illusion in which the length of the gesture used to produce a sound alters the perception of the sound’s duration. This contradicts the widely accepted claim that the auditory system dominates in temporal tasks (such as estimating duration) because of the superior quality of its temporal information (the optimal integration hypothesis). In the first of four experiments we show that impact gestures influence the duration rating of percussive sounds but not of sustained ones. In the second experiment we show that the illusion is present even if the percussive sound occurs up to 700 ms after the visible impact, and that it disappears if the percussive sound precedes the visible impact. In the third experiment we show that only the motion after the visible impact influences the perceived tone duration. The fourth experiment — in which we replaced the video gestures with the text “Long” and “Short” — shows that the phenomenon is not due to response bias. We conclude that causality plays an important role in sensory integration.

Keywords: sensory integration, causality, auditory, visual, optimal integration

Schutz and Lipscomb (2007) discovered an illusion in which a visual event affects the perceived duration of an accompanying sound. They made several videotapes of an outstanding performer playing single notes at various pitches on the marimba. The performer — who believed that he could control the duration of notes by varying the movement of the mallet — performed each note in two ways. We will call the gesture with which he tried to produce long notes the long gesture and the gesture with which he tried to produce short notes the short gesture. When Schutz and Lipscomb asked participants to judge the durations of sounds produced with these two gestures in the absence of visual information, they judged the durations of the notes to be equal. However, when they heard the sounds while watching the video, they judged notes produced by long gestures to be longer than notes produced by short gestures.

This finding is at odds with the consensus view that in temporal tasks, audition influences vision (Shimojo et al., 2001) whereas audition is not influenced by vision (Welch, DuttonHurt, & Warren, 1986; Welch & Warren, 1980; Walker & Scott, 1981; Shipley, 1964; Shams, Kamitani, & Shimojo, 2002). The perception of event duration (which concerns us here) is no exception (Welch & Warren, 1986). Here we show that this view is not valid when the relationship between the auditory and the visual information is causal (such as when a visible impact produces a percussive sound).

The consensus view is based on a large body of research that supports the optimal integration hypothesis, according to which intermodal conflicts are resolved by giving more weight to the modality that provides the more reliable information (Ernst & Banks, 2002; Alais & Burr, 2004): when an audio-visual conflict is spatial, vision dominates because the spatial acuity of the visual system is better; when it is temporal, audition dominates because the temporal acuity of the auditory system is better. (A special case of this hypothesis is the modality appropriateness hypothesis, according to which dominance is a function of modality — Welch et al., 1986; Welch & Warren, 1980.)

Examples of the applicability of the optimal integration hypothesis abound. The superior spatial acuity of vision accounts for the ventriloquism effect, in which speech appears to originate from the moving lips of a puppet (Jack & Thurlow, 1973) and its non-speech analogs (Jackson, 1953; Thomas, 1941; Witkin, Wapner, & Leventhal, 1952; Bertelson & Radeau, 1981; Bertelson, Vroomen, de Gelder, & Driver, 2000). Likewise, the superior temporal acuity of audition can account for cases in which auditory stimulation affects visual perception, but not vice versa: (a) in simultaneous flash-tone pairs, ratings of flash duration are affected by the tones (Walker & Scott, 1981); (b) the perceived number of visual flashes is affected by the number of tones presented (Shams et al., 2002); (c) the perceived rate of visual flicker is affect by the rate of auditory flutter (Shipley, 1964; Welch et al., 1986); (d) estimating the temporal order of flashes is affected by the temporal order of tones (Fendrich & Corballis, 2001); and finally, (e) hearing temporally discrepant auditory and visual stimuli affects the subsequent visual perception of

Correspondence should be addressed to Michael Schutz or Michael Kubovy, P.O.Box 400400, Department of Psychology, University of Virginia, Charlottesville, VA 22904–4400; e-mail: schutz@virginia.edu or kubovy@virginia.edu. This research was supported by a grant from NIDCD R01 DC 005636 (Michael Kubovy, PI), and was performed in partial fulfillment of the requirements for an MA degree. We thank S. Fitch of Mustard Seed Software for the outstanding programs we used to run the experiments and J. Armontrout for his devoted help in both producing stimuli and running participants. J. Armontrout, S. Boker, W. Epstein, R. Keen, T. Salthouse, J. Shatin, M. Yu gave us helpful comments on various versions of the manuscript.
temporal rate (Recanzone, 2003).

Even though in general the optimal integration hypothesis predicts that one modality will be given more weight in each type of intersensory conflict (which is essentially the claim of the modality appropriateness hypothesis), it also correctly predicts that by artificially degrading the input to the dominant modality one can reverse its dominance (which the modality appropriateness hypothesis cannot). For instance, when Wada, Kitagawa, and Noguchi (2003) paired fluttering tones with flickering lights, if the auditory information was unambiguous, the expected audition—vision effect occurred, but when it was degraded, the reverse was true. (Similar effects have been reported by Battaglia, Jacobs, & Aslin, 2003; Alais & Burr, 2004.)

Causality and cross-modal integration

Optimal integration cannot tell the whole story of cross-modal integration because it focuses on predicting the winner of a competition. However, integration is generally the outcome of a collaboration rather than a competition. When we walk down a street hear many sounds such as barking, vehicle noises, and voices. We also see many things, such as dogs, vehicles, and people. The role of perception is not to optimally integrate these sounds and sights based on information quality, but to integrate only those that specify a common event, such as the dog with its bark. Therefore, cross-modal competition can take place only when the conflicting information in each modality is deemed to have been produced by the same source. Listening to background music might affect your reading of this article because your attentional resources are limited or because there may be leakage between modalities, but this would not be a case of the cross-modal system trying to use the music and the text to identify the source of the current audio-visual stimulation. Unless a system has determined that a single object or event is there to be perceived, say, an audible and visible object — there can neither be competition nor cooperation between the modalities.

This requirement has been called the “unity assumption” (Welch, 1972; see also: Spence, 2007; Welch & Warren, 1980; Welch, 1999; Vatakis & Spence, 2008; Vroomen, 1999) as well as the “identity decision” (Bedford, 2001a, 2001b, 2004) — whenever the congruence of two sensory inputs is great, they will be perceived to have been caused by a single distal event, and therefore cross-modal binding will more readily occur. We assume that cross-modal systems engage in a search for identity cues to determine whether the unity assumption is justified and whether an identity decision should be reached before resolving any cross-modal conflict.

What then does it do for bimodal stimulation to trigger an identity decision? Four types of identity cues have been reported:

1. Voice-face gender congruence. For example, according to Vatakis and Spence (2007) when a visible face speaks, gender congruence is an identity cue. When this identity cue is present, Vatakis and Spence assume that the face and the voice are strongly bound, which renders the stimulus resistant to analysis. As they expected, they found that it is easier to detect which came first, a speech sound or lip movements, when the two sources provide a weak identity cue — when a female voice emanates from a male face — than when they provide a stronger one — when voice and face are female.

This identity cue does not seem to be triggered under all conditions. For example, the McGurk effect (McGurk & MacDonald, 1976, in which visible lip movements alter a listener’s perception of spoken syllables) is insensitive to such discrepancies (Green, Stevens, Kuhl, & Meltzoff, 1991), as long as the lip movements and the syllables are fairly well synchronized.

2. Synchrony. In the McGurk effect, when the two sources are synchronized to within 180 ms (Munhall, Gribble, Sacco, & Ward, 1996) the two sources appear to trigger an identity decision, and the phenomenon is robust: (a) it is unaffected by manipulations of word meaning or sentence context (Sams, Manninen, & Surakka, 1998), (b) it is insensitive to discrepancy between the gender of the face and the voice (Green et al., 1991), and (c) it requires only a minimum of acoustic information (Remez, Fellowes, & Pisoni, 1998). However, beyond 180 ms, the worse the lip–speech synchronization, the weaker the effect (Munhall et al., 1996).

3. Spatial congruence. In the ventriloquist illusion, the greater the spatial distance between the origin of the sound and visual information, the weaker the effect (Jack & Thurlow, 1973).

4. Affective congruence. de Gelder, Pourtois, and Weiskrantz (2002) studied audiovisual integration in the processing of speech in two patients with blindsight. They presented these patients with two types of audiovisual stimuli: naturalistic pairs (in which an emotional voice accompanied an emotional face) and semantic pairs (in which an emotional voice accompanied an emotional scene) in two combinations: emotionally congruent and emotionally incongruent. They asked the patients to identify the gender of the voice, while recording EEG. Under free-viewing visual presentation, incongruence decreased the amplitude of auditory-event-related potentials for both types of pairs. But when they presented the visual stimulus to a patient’s blind spot, the effect occurred only for naturalistic pairs. So it appears that nonstriate circuitry cannot detect identity cues for semantic pairs, whereas striate vision can.

For all their heterogeneity, these manifestations of identity cues share one feature — they were all obtained with speech stimuli. This raises the question whether the identity decision applies to non-speech stimuli. Experiments by Vatakis and Spence (2008) provide an answer. They presented video clips of object action and musical notes which were congruent (e.g., the sight of a piano key being struck accompanied by the corresponding sound) or not (e.g., the sight of the piano accompanied by a guitar pluck). They asked participants to report which modality stream had been presented first. Their results (as well as Radeau and Bertelson’s, 1977) implied that the unity assumption does not apply to non-speech stimuli.

Vatakis and Spence’s conclusions — that the unity as-
sumption applies only to speech stimuli — leave us at a loss to account for the Schütz–Lipscomb illusion. And yet there may be a way out, if we recall that humans readily detect causal relationships within and across modalities (Fisher, 1962; Guskis & Troje, 2003).

The most conspicuous example of cross-modal causality is the perception of a causal link between a visible impact and a percussive sound. It has been investigated by Arrighi, Alais, and Burr (2006), who created drumming videos in which they desynchronized the sound track and the video by different degrees, and found that participants perceived maximal synchrony when the audio lagged slightly behind the video. When they replaced the video of the drummer with dots whose movements matched the drummer’s, the lags required for perceived synchrony were the same. But when the dots oscillated at constant speed, the required lags were considerably smaller.

Cross-modal causality may also play a role in Sekuler, Sekuler, and Lau’s (1997) discovery of the effect of sound on the perception of bouncing: two circles approach each other, overlap briefly and then continue on their respective paths. This display can be seen as balls colliding and recoiling or passing through each other. A tone played at the moment of contact increased the likelihood that the event is seen as a bounce. Watanabe and Shimojo (2001; see also Shimojo et al., 2001) showed that when the sound that coincided with moment of contact was preceded and followed by identical tones (300 ms before and after the moment of contact), the effect disappeared (this is because auditory grouping occurs before intersensory pairing, as Keetels, Stekelenburg, & Vroomen, 2007, have shown). When the pitch of these contextual sounds differed from the pitch of the critical tone, the effect was restored.

The work of Stekelenburg and Vroomen (2007) reinforces the likelihood that cross-modal causality plays a central role in cross-modal integration. Using ERP as their response variable, they compared the cross-modal integration of audio-visual speech to non-speech pairings such as an audio-visual clap or an audio-visual clink of a spoon against a cup. They found that (a) the effects they observed were larger for the impact/percussion pairs than for the audio-visual speech pairs, (b) incongruity did not decrease the magnitude of the effect with impact/percussion pairs, (c) neither audio-visual tearing of paper nor sawing of wood had an effect. They conjectured that the latter effect is due to the absence of visible anticipatory motion, an idea we will revisit in the General Discussion.

All this suggests a binding by causality hypothesis: percussive sounds have a propensity to bind with the visible movements that caused them, which we test in Experiment 1. We will use the Schütz–Lipscomb illusion to gauge this binding: we expect the illusion not to occur when visible impacts are paired with sustained sounds. The experiment will also address two alternative hypotheses:

(a) The uncertainty hypothesis. If the participants were (i) more uncertain about the duration of percussive sounds than the duration of the gestures, and (ii) less uncertain about the duration of the sustained sounds than the duration of the gestures, then the optimal integration hypothesis could predict the illusion: in the first case the visual impact would affect the perceived duration of the sound; in the second, it would not. We therefore included an audio-alone condition that will tell us whether, in the absence of a visual information, duration ratings of percussive sounds are more variable than duration ratings of sustained ones. We examine this hypothesis in Experiment 1.

(b) The response-bias hypothesis. If the gestures, by being suggestive of greater or lesser duration, affected the ratings but not the perceived durations, we would say that the illusion is due to response bias. (That cross-modal influences my involve bias is demonstrated by Arieh & Marks, 2008, who have shown that response bias is involved in an effect of hearing on visual selective attention.) One way to undermine this hypothesis is to show that under some conditions, the suggestive gestures have no effect, which rules out response bias. We will tackle this hypothesis head-on in Experiment 4.

Experiment 1: Percussive vs. Sustained Sounds

In this experiment we ask whether the illusion, hitherto obtained with marimba tones, occurs with sustained sounds. If binding by causality plays a role in cross-modal integration, a gesture should only influence sounds that it could have produced: impact gestures should affect the perception of percussive sounds but not of sustained ones.

Method

The experiment took place in a quiet room using an Apple Macintosh G4 computer running the MAX/MSP program which controlled stimulus presentation. Stimuli were presented on a ViewSonic E790B monitor (resolution: 1280×1024; refresh rate: 85 Hz) and Sennheiser HD 580 Precision headphones. Participants could adjust loudness during the warm-up period.

Stimuli. Each stimulus had an auditory and a visual component:

(a) Auditory component: We used two percussive (piano, marimba) and four sustained (clarinet, french horn, voice, and white noise) sounds. As Figure 1 shows, percussive sounds begin with a sharp attack followed by an exponential decay. We used two versions (long and short) of each timbre. Their durations were approximately those of two of the marimba tones used by Schütz and Lipscomb: E1 (~ 82 Hz) and D4 (~ 587 Hz).

(b) Visual component: We derived the visual stimuli for the gestures from Schütz and Lipscomb, depicting a marimba performing single notes with long or short gestures. As Figure 2 shows, the videos included the performer’s head and torso in addition to the hand and arm motion involved in the impact. In the audio-alone conditions the visual component was a blank screen.
Figure 1. Amplitude envelopes of timbres used in Experiment 1, each in two versions, long (top) and short (bottom). Percussive sounds such as the piano and marimba (right) exhibit a sharp attack, followed by a gradual decay. Sustained sounds (left) exhibit a slower attack and a sustained period between attack and decay.

Figure 2. The videos showed the upper body of the marimbist, including full stroke preparation and release. Reprinted from Schutz & Lipscomb (2007) with permission from Pion Limited, London.

Conditions. We presented stimuli in two conditions:
(a) Audio-visual. We crossed the twelve sounds (6 timbres with 2 levels of duration) with the two visible gestures, for a total of 24 audio-visual stimuli.
(b) Audio-alone. The twelve sounds were also presented alone.

Participants and Procedure. We recruited twenty-six participants from introductory courses in psychology: they participated for course credit. They were told that some of the stimuli contained mismatched auditory and visual components, and were asked to judge the duration of the tone independent of the visual information with which it was paired. Stimuli were presented six times within six blocks (one presentation per block) under two conditions: (a) as audio-visual stimuli combining the visible gesture and sound, and (b) as audio-alone. Blocks and stimuli within blocks were randomized.

Participants rated the duration of the sounds by using an unmarked 101-point slider with endpoints “Short” and “Long.” To ensure that they were attending to the visual information they also rated the degree to which the auditory and visual components of the stimulus agreed, by using a second slider with endpoints “Low agreement” and “High agreement.” Although one might worry that this secondary task might interfere with the primary task, Rosenblum and colleagues (Rosenblum & Fowler, 1991; Saldana & Rosenblum, 1993) have shown that it does not. Since the purpose of these ratings was only to draw the participants’ attention to the visual component of the stimulus, we did not analyze them.

Results

Data analyses. Our conclusions are based on linear mixed-effects models (also known as multilevel analyses or hierarchical linear models) estimated by restricted maximum likelihood (REML), using the function lmer (Bates & Sarkar, 2007), running on R (Ihaka & Gentleman, 1996). Several textbooks (Baayen, 2008; Kreft & Leeuw, 1998; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999) present mixed-effects analyses, which have considerable advantages over traditional so-called repeated-measures analyses based on quasi-F tests, by-subjects analyses, combined by-subjects and by-items analyses, and random regression (Baayen, Davidson, & Bates, in press and Maxwell & Delaney, 2004, Part IV). For each set of data, we obtain estimates of effects from a minimal adequate (or reduced) model, which is (a) simpler than the maximal model (which contains all factors, interactions and covariates that might be of any interest), (b) does not have less explanatory power than the maximal model, (c) has no submodel that is deemed adequate. The minimal adequate model is obtained from the maximal model by a process of term deletion (also known as backward selection; for an introduction, see Crawley, 2007, pp. 323–329). We report each result in terms of an effect (and its standard error, se, in parentheses), from which a Cohen effect size, d, can be obtained by dividing the effect by its se. To these we add a 95% confidence interval (henceforth ci), as well as a p-value for a test of the null hypothesis that the effect in question is 0. By presenting the correct error
bars for mixed models we follow the recommendations of Loftus (2002, with appropriate allowance for the differences in statistical techniques); and by minimizing the role of null-hypothesis statistical tests, we implement the recommendations of the APA Task Force on Statistical Inference (Wilkinson, 1999).

The binding by causality hypothesis. Figure 3 shows that gesture length significantly affected duration ratings for three out of four percussive sounds but not for the sustained ones (clarinet, voice, french horn, and white noise), and the effects on the four percussive sounds were larger than the effects on the sustained sounds. This is in line with binding by causality. We first determined the likelihood that the effect on the four percussive sounds would be higher than the effect on the other eight sounds. To do this we assumed that the twelve effects were normally and identically distributed (i.e., no differential effect on percussive sounds) and determined by simulation (using the extreme-value distribution) that a particular set of four means would be higher than the other eight with a probability of $p = 0.002$. The average effect of gesture on the marimba sounds was $7.5 \pm 3.1$ points (95% confidence interval, $c$: $[1.4, 14], p = 0.01$) higher than their average effect on the piano sounds. The average effect of gesture on percussive sounds was $6.9 \pm 2.5$ points higher than their effect on sustained sounds ($95\% \ c$: $[2.4, 11], p = 0.004$).

The uncertainty hypothesis. Uncertainty about a stimulus feature is measured by the variability of observer responses (Ernst & Banks, 2002; Alais & Burr, 2004). If the uncertainty hypothesis were true, the variability of duration ratings of percussive sounds in the audio-alone condition would predict the audio-visual data. To find out, we analyzed the ratings of the twelve audio-alone stimuli (shown in Figure 4, and for each we measured the variability of audio-alone ratings by taking $\sqrt{\text{standardized residuals}}$ (a standard measure for the assessment heteroscedasticity, see Cleveland, 1993, p. 105). Figure 5 plots the effects of gesture on duration ratings (which are the same as in Figure 4) as a function of the variability of audio-alone ratings. If the uncertainty hypothesis were true, then: (a) in the audio-alone condition, percussive sounds would be more variable than sustained sounds. This is not the case: percussive sounds are only $0.02 \pm 0.03$ points more variable than sustained sounds ($95\% \ c$: $[-0.04, 0.07], p = 0.5$). (b) Sustained sounds would cluster in lower left quadrant of Figure 3, and the percussive sounds would cluster in the upper right quadrant. They do not: two of the percussive sounds are above the median variability, and two are below. (c) Finally, a regression of the audio-visual data on the variability of duration ratings of percussive sounds in the audio-alone condition gave $R^2 = 0.09; R^2_{adj} \approx 0$. Another way to reach the same conclusion: If the true value of this effect were 0 for these eight sounds, we would expect only half of these to be positive; under this expectation, the probability that five or more are positive (i.e., three or fewer are negative) is $p = \frac{8!}{5!3!} \left(\frac{1}{2}\right)^8 = 0.36$. In other words, we find no support for the uncertainty hypothesis.

The response-bias hypothesis. If the gestures merely affected the ratings by suggestion, then they would have had the same effect on the sustained sounds as on the percussive sounds. That the effect on sustained sounds is negligible shows that the difference in the magnitude of the illusion for the two types of sound cannot be due to response bias.

The effect of perceived sound duration on the illusion. To answer this question, we computed a Kendall rank correlation ($\tau$) between the effect of gesture in the audio-visual condition (Figure 3) and the corresponding mean ratings of sound duration the audio-alone condition (Figure 4). The wide range of the latter (from 7.5 to 34.5) reassures us that a small value of $\tau$ is not due to restriction of range. The correlation was $\tau = 0.42$ which, for $H_0$ that the two orders are independent, gives $p = 0.06$. This is an indication that sound duration may have a effect on the illusion, but the evidence is not strong enough to draw a conclusion.

The effect of video on the sensitivity of duration ratings. Having established that gesture type affects duration ratings, we wondered what effect the presence of video might have had on them. To answer this question, we compared ratings of duration in the audio-visual trials with ratings of duration in the audio-alone trials. These audio-alone ratings ranged widely, as we saw in Figure 4. What then is the functional relation between audio-visual and audio-alone ratings? The results, summarized in Figure 6, show that the slope of the linear function relating the audio-visual to the audio-alone trials is marginally less than $1$: $0.88 \pm 0.08$ points (95% $c$:
The effect of gesture

\[
\begin{array}{l}
\text{Clarinet2} \\
\text{Whitenoise1} \\
\text{Horn1} \\
\text{Voice1} \\
\text{Voice2} \\
\text{Whitenoise2} \\
\text{Clarinet1} \\
\text{Horn2} \\
\text{Piano2} \\
\text{Piano1} \\
\text{Marimba2} \\
\text{Marimba1} \\
\end{array}
\]

Figure 3. Experiment 1: Degree of visual influence on twelve sounds, as measured by the difference between their duration ratings when paired with long and short gestures. The gestures exerted a strong influence on the marimba sound, a moderate influence on the percussive piano sound, and no influence on the perception of sustained sounds of the clarinet, voice, french horn, and white noise. Error bars represent 95% cIs. (See Figure 6 for a somewhat different analysis.)

\[
[0.73, 1.03], p \approx 0.05.
\]

This suggests that the presence of video may reduce the sensitivity of participants to the durations of the sounds.

**Discussion**

Experiment 1 produced two principal results: (a) The Schutz and Lipscomb illusion does not occur with sustained sounds, only with percussive ones. (b) The illusion is stronger when the visible impact matches the timbre of the percussive sound more closely (i.e., the effect of the video on marimba sounds is greater than its effect on piano sounds).

In light of these results, we can update the hypothesis of binding by causality. The refined hypothesis stipulates that when a visible event and a sound occur in temporal proximity, the pairing gives rise to an impression of causality, \( C(\text{visible event} \rightarrow \text{sound}) \), and that such impressions can be ordered. If this is the case, we can safely assume that \( C(\text{visible impact} \rightarrow \text{marimba}) > C(\text{visible impact} \rightarrow \text{piano}) > C(\text{visible impact} \rightarrow \text{sustained sound}) \approx 0 \) (meaning that no impression of causality links a visible impact with a sustained sound). If the magnitude of the illusion depends on the impression of causality, we would expect the results we obtained: the largest illusion with marimba sounds, a weaker illusion with piano sounds, and no illusion with sustained sounds.

We ruled out the uncertainty hypothesis by showing that the variability of ratings for percussive and sustained sounds did not differ in the audio-alone condition, and did not predict the effects we obtained in the audio-visual condition. Likewise, we ruled out the response-bias hypothesis by showing that the effect of gesture on sustained sounds is negligible. Finally, we noted that the slope of the linear function relating audio-visual and audio-alone trials may be less than 1.0, which — if confirmed in other experiments — would be evidence that dividing the participants’ attention between the auditory and visual sources degrades their ability to discriminate among sounds.

**Experiment 2: The effect of asynchrony**

If the hypothesis of binding by causality is correct, then any disruption of the temporal ordering of cause and effect should eliminate the Schutz–Lipscomb illusion. In this experiment we manipulate this order: the visible impact was either synchronous with the sound, preceded it, or followed it.
Figure 5. Evidence that uncertainty did not affect the results of Experiment 1. Duration ratings for the twelve sounds for the audiovisual trials as a function of the variability of duration ratings in the audio-alone trials. The vertical dotted line passes through the median variability, and the horizontal dotted line passes the median effect. If uncertainty affected the illusion, the data for the percussive sounds would cluster in quadrant I (upper right) and the data for the sustained sounds would cluster in quadrant III (lower left). The SE bar is the average of the standard errors used to determine the cts in Figure 3.

Method

The experiment was identical to Experiment 1, except in the ways we describe next.

Stimuli. The stimuli were derived from the marimba videos and the marimba sound tracks used in Experiment 1.

(a) Auditory component: We used marimba tones of different durations, which we controlled by manipulating: (1) Sound termination: by using either natural tones that decayed naturally, or damped tones that were manually damped soon after the bar was struck. (2) Natural decay time: because bars tuned to lower frequencies ring longer (Bork, 1995), we varied the frequencies of the marimba sounds. We crossed the two types of sound termination (damped, natural) with three musical pitches: E1 (~82 Hz), D4 (~587 Hz), and G5 (~1568 Hz).

(b) Visual component: The original long and short gestures from the previous experiment served as the visual stimuli (Figure 2).

The sound was (a) synchronous with the visible impact, (b) preceded it by 400 or 700 ms, or (c) followed it by 400 or 700 ms. Thus the levels of offset were: −700, −400, 0, 400, 700 (negative values denote the audio-alone duration ratings).

Figure 6. Duration ratings for the twelve sounds for the audiovisual trials in Experiment 1 as a function of ratings of identical sounds in the audio-alone condition, by timbre class (sustained vs. percussive) and visible gesture (long vs. short). Error bars represent ±1 SE. The lines represent the best-fitting linear regression functions.

Figure 7. Natural (top row) and damped (bottom row) marimba tones used as auditory stimuli for Experiment 2.

Conditions. We presented the sounds alone or with the video.

Participants and Procedure. We recruited ten new participants, and paid them $10 for their participation. They went through 264 trials, in five randomized blocks (four audiovisual blocks of 60 trials and one audio-alone block of 24 trials).

1 These timings are accurate within a 33 ms window.
Results

The binding by causality hypothesis. As Figure 8 shows, the illusion was absent in the audio-first conditions: the average effect in the audio-first conditions was 7.2 (±4.4) points (95% cr: [−1.1, 16.3], p = 0.84). It was present in the synchrony and the video-first conditions. The effect in the synchrony condition was larger than in the video-first conditions: the difference between the effect at synchrony and the average effect in the video-first conditions was 10.8 (±2.1) points (95% cr: [6.8, 14.9], p ≈ 0).

The uncertainty hypothesis. As in Experiment 1, the sounds varied widely in perceived duration in the audio-alone condition (Figure 9). To determine whether the variability of these duration ratings predicts the audio-visual data, we plotted the effect of gesture as a function this variability (Figure 10). As is evident from the trend line (labeled “mean”), the magnitude of the illusion in the audio-visual condition does not increase substantially as a function of the variability of audio-alone ratings. (We note that the range of the measures of variability was similar to the range for Experiment 1: 0.74 to 0.95).

The response-bias hypothesis. Our analysis here is parallel to the one for Experiment 1, in which we found no evidence for response bias. Here, however, even though we found no significant effect in the audio-first conditions (−400, −700 in Figure 8), we do have some evidence of response bias. If the true value of this effect were 0 for these twelve sounds, we would expect only half of them to be positive in our data; and yet the effect of gesture was positive.
for the twelve audio-first conditions, an extremely unlikely event: $p = 0.512 = 0.00024$.

**Effect of perceived sound duration on the illusion.** We wish to ascertain whether the order of effects for the six sounds (Figure 8) was consistent across offsets. (As Figure 9 shows, the mean perceived durations of these sounds varied widely: from 25 to 87, a range of 62, compared to a range of 67 in Experiment 1.) To this end we computed the intra-class correlation (icc, often used to determine inter-rater reliability, Shrout & Fleiss, 1979) among the orders of effect sizes across offsets (Falissard, 2008). We obtained an icc = -0.12 (95% cr: [-0.21, -0.0001], $p = 0.05$, computed by bootstrap simulation); this is a mild indication of disagreement, which is of no concern. Only if the agreement were positive would we have grounds to conclude that perceived sound duration had an effect on the illusion.

**The effect of video on duration ratings.** Having established that audio-visual asynchrony affects the illusion, we wondered — as we did in Experiment 1 — what effect the presence of video might have had on duration ratings. To answer this question, we compared ratings of duration in the audio-visual trials with ratings of duration in the audio-alone trials. These audio-alone ratings ranged widely, as we saw in Figure 9. What then is the functional relation between audio-visual and audio-alone ratings? The results, summarized in Figure 11, show that the slopes of the linear functions relating the audio-visual to the audio-alone trials are less than 1 (although not necessarily significantly so): 0.92 (95% cr: [0.55, 1.3]), 0.95 (95% cr: [0.58, 1.3]), 0.72 (95% cr: [0.46, 0.99]), 0.81 (95% cr: [0.46, 1.2]), 0.82 (95% cr: [0.43, 1.2])). It is noteworthy that the offset for which the effect of gesture is largest (offset = 0), the slope is significantly lower than 1 (albeit marginally so). As in Experiment 1, this suggests that here too participants were less sensitive to differences in sound duration in the presence of a video. In addition it hints to the intriguing possibility that this decrease in sensitivity is modulated by the quality of the link between the visual and the auditory information.

**Discussion**

Experiment 2 produced three principal results: (a) The Schutz and Lipscomb illusion does not occur when the percussive sound precedes the visible impact, only when the sound follows the visible impact. (b) The illusion is stronger when the percussive sound and the visible impact are simultaneous than when the sound follows the impact. (c) The illusion is still present when sound comes 400 or 700 ms after the visible impact.

The literature would lead us to expect the first two results: auditory-visual cross-modal effects are generally weaker when the sound precedes the visual information than when it is simultaneous with the visual information or follows it (in the McGurk effect by Slutzky & Recanzone, 2001; Munhall et al., 1996, in the audiovisual perception of drumming by Arrighi et al., 2006, in temporal ventriloquism by Stekelenburg & Vroomen, 2007) The third result is unexpected because the cross-modal effects just cited do not persist beyond a delay of 200 ms. This suggests that a stronger form of cross-modal binding may be occurring in this situation. It should be noted, however, that a delay of 700 ms is not ecologically impossible. Indeed at 340 m/s, sound can travel 238 m in 700 ms, 2.2 or 2.3 times a football or a soccer field. Under free-field conditions, a loud sound would be audible at that distance (say the loudness of a pneumatic hammer 1 m away, roughly 106 dB-SPL, will have dropped to the loudness of a quiet restaurant, roughly 48 dB-SPL at 256 m).

Here too we have evidence against the uncertainty hypothesis: as Figure 7 shows, decay times for individual tones varied considerably as a function of pitch level and production type (natural vs. damped), and as a result they were judged to vary considerably in duration (Figure 9). Nevertheless, the magnitude of the illusion was independent of these variations. We also showed that the uncertainty in the rating of the durations of these sounds was unrelated to the magnitude of the illusion (Figure 10).

Finally, let us look at our evidence regarding the response-bias hypothesis. We noted that the twelve effects obtained for the negative offsets were positive, and that the probability of such an event is negligible. Two accounts of this finding can be proposed: that (a) there is a small visual effect on auditory judgments even when the sound precedes its visible cause, or that (b) the video produces an 8-point response bias on the rating of the duration of the sound that preceded it. Our data do not offer a resolution.

Whichever of these accounts may be the case, the main findings of this experiment support the binding by causality hypothesis: the illusion is contingent upon causality. The effect occurred only when there was a plausible visual-auditory causal link. Furthermore, as with Experiment 1, the effect was graded: it was strongest when the causal link was most plausible (synchrony), weaker when the causal link was less likely (audio-lag), and very weak when the causal link was impossible (audio-lead).

**Experiment 3: Which part of the gesture is responsible for the illusion?**

In Experiment 3 we determine which portion of the gesture (pre- or post-impact) is responsible for the illusion.

**Method**

The experiment was identical to Experiments 1 and 2, except in the ways we describe next.

**Stimuli**

We used the auditory and visual stimuli from the 0 ms offset of Experiment 2 create a segment (pre-impact) video that shows the gesture prior to the impact and freezes when the sound begins, and a segment (post-impact) video that starts frozen on the moment of impact until the sound begins and displays the post-impact gesture. A segment (both) stimulus consisted of the original videos with the complete gesture.
Participants and Procedure

Twenty-nine new participants from introductory courses in psychology received course credit for their participation. Each went through 324 trials in 18 blocks: six audio-visual blocks (36 trials each), six audio-alone blocks (6 trials each), and six video-alone blocks (12 trials each), in which both block and trial order randomized. During the video-alone condition (used only in this experiment), participants were asked to rate the relative duration of the gesture on the same scale used for the audio-visual and audio-alone presentations.

Results

The left panel of Figure 12 summarizes the results of the main part of the experiment, the audio-visual condition. It shows that the visual influence is due to the post-impact segment: gesture affected ratings only when it was visible concurrently with the sound (in both the pre- and post-impact videos). The right panel summarizes the results of the video-alone condition. It shows that the participants could tell which gesture was long and which was short from the information in both segments, although the difference was more vivid when the post-impact segment was visible.

The importance of the post-impact portion of the gesture. In the post-impact segment in the audio-visual condition, the gesture exerted a 12 (±2.1) point influence on duration ratings (95% cr: [8.5, 16.5], p < 0.01). This is only slightly less than its influence when both segments (the whole gesture) were visible: a 14 (±2) point effect. In contrast, the pre-impact segment showed a negligible 3.1 (±2.1) point effect (95% cr: [-1.0, 7.1], p = 0.13). This lack of influence does not reflect visual ambiguity, as illustrated by the 21.9 (±5.5) point difference between ratings of the pre-impact gestures when presented as video-alone (95% cr: [11, 33], p < 0.01).

The difference between segments. The effect of the post-impact portion of the gesture was 9.4 (±1.6) points larger (95% cr: [6.3, 12.5], p < 0.01) than the effect of pre-impact portion of the gesture. In contrast, in the both condition,
the gesture had a negligibly larger effect than in the post-impact condition: a 1.7 (±1.6) point difference (95% ci: [-1.5, 4.9], \( p = 0.3 \)), which once more shows that the post-impact segment of the gesture is responsible for the visual influence.

The gestures were discriminable in the video-alone condition. As the right panel of Figure 12 shows, the participants discriminated well between long and short gestures in the video-alone condition. There was a 49 (±5.5) point effect of gesture length (95% ci: [38, 59], \( p \approx 0 \)) in the post-impact segment and a slightly larger 53 (±5.5) point difference (95% ci: [42, 64], \( p \approx 0 \)) when both segments were visible. The effect for the pre-impact segment was smaller, 21.9 (±5.5) points (95% ci: [11, 33], \( p \approx 0 \)), but still larger than the largest audio-visual influences in this experiment.

The uncertainty hypothesis. As in the previous experiments, the ratings of the durations of the sounds in the audio-alone condition varied widely (Figure 13). Figure 14 shows the effect of gesture as a function of the variability of these ratings. It is evident from the trend line (labeled “mean”) that the magnitude of the illusion in the audio-visual condition does not increase as a function of the variability of audio-alone ratings. (We note that the range of the measures of variability was similar to the range for the previous experiments: 0.73 to 0.94).

The response-bias hypothesis. As in Experiment 2, although we found no significant effect in the pre-impact condition (Figure 12), we do have some evidence of response bias. If the true value of this effect were 0 for these six sounds, we would expect only half of them to be positive in our data; that the effect of gesture was positive for these six conditions is unlikely: \( p = 0.5^8 = 0.016 \).

Effect of perceived sound duration on the illusion. As in Experiment 2, we tested whether the order of effects for the six sounds (Figure 12) was consistent across video segments. (As Figure 13 shows, the mean perceived durations of these sounds varied widely: from 23 to 78, a range of 55, compared to 67 and 62 in the first two experiments.) We computed the intra-class correlation and obtained an ICC = 0.6 (95% ci: [-0.2, 0.8], \( p \approx 0.05 \), computed by bootstrap simulation); there is no evidence of agreement, and we conclude that perceived sound duration did not have an effect on the illusion.

The effect of video on duration ratings. In this experiment too, audio-alone ratings ranged widely, as we saw in Figure 13. The functional relation between audio-visual and audio-alone is summarized in Figure 15. It shows that the slope of the linear function relating the audio-visual to the audio-alone trials is definitely less than 1: 0.78 (±0.02) points (95% ci: [0.76, 0.81], \( p > 0.05 \)). This suggests that the presence of video reduced the sensitivity of participants to the durations of the sounds.

Discussion

We have identified a major source of the illusion: the visual information concurrent with the sound, the post-impact gesture. Two observations support this claim (a) The effect of gesture is just about the same when the entire gesture is visible as when only the post-impact segment is visible. (b) The effect of the pre-impact gesture is much smaller than the effect the post-impact gesture. It would appear that in an
event in which a visual cause (the pre-impact segment and the impact) producing both a visual effect (the post-impact segment) and an auditory one (the percussive sound).

A caveat: We can conceive of the possibility that there was no difference between the effect of the whole gesture and the effect of the post-impact gesture because participants were imagining the pre-impact gesture in the trials with only post-impact motion. It would be interesting to compare the results we obtained here with those we could obtain with participants who had never seen the pre-impact portion of the stroke before going through trials in which only the post-impact segment was shown. Nevertheless our conclusion would be essentially unchanged even if the the effect of gesture was somewhat smaller than the effect of the whole ges-
ture, as long as it remained substantially larger than the effect of gesture in the pre-impact condition.

We can again update the hypothesis of binding by causality. When a visible event causes a visible and an audible effect, the two effects are bound and they can influence each other (although in our case the influence is uni-directional, we cannot rule out that under other circumstances, the direction of influence might be reversed).

Again we have evidence against the uncertainty hypothesis: the uncertainty in the rating of the durations of the sounds in the audio-alone condition was unrelated to the magnitude of the illusion in the audio-visual condition (Figure 14).

Finally we revisit the response-bias hypothesis. We noted that the six effects obtained for the pre-impact condition were positive, and that the probability of such an event is small. Just as in Experiment 2, we can propose two accounts of this finding: that (a) the visible difference between long and short gestures had an effect on the duration ratings, or that (b) this information produces a 3-point response bias on the rating of the duration of the sound that preceded it. Our data do not allow us to adjudicate between the two accounts.

**Experiment 4: A test of the response-bias hypothesis**

Even though the three experiments described so far offer compelling evidence against the response-bias hypothesis, we designed Experiment 4 as a final test of this hypothesis. We compared the effect of the videotaped gesture to the effect of a suggestive text — the written words long and short. Assuming that text cannot alter the perception of concurrent auditory information, if we found an influence of text, we would suspect that top-down influences, such as bias, can play a role in the illusion.

**Method**

This experiment followed the methodology used in the first three experiments.

**Stimuli**

We used the same stimuli as in the 0 offset condition of Experiment 2. We created the text condition by replacing the long and short gestures with the words “Long” and “Short.”

1. **Auditory component.** The sounds were the same as in Experiment 2: two types of marimba tones (natural, damped) performed at three pitches: E1 (~82 Hz), D4 (~587 Hz), and G5 (~1568 Hz).

2. **Visual component.** There were two visual conditions:

   a. **display (video):** The long and short gestures used in Experiment 2.

   b. **display (text):** We replaced the videos of gestures with the text “Long” and “Short” written in black on white. The text was visible for the same duration as the downstroke [MK: or was upstroke correct?] in the videos. This maximized the chance of a visual influence: the marimba tones began approximately 1 – 1.5 seconds after the text, giving participants the time to read it before hearing the sound.

**Participants and procedure**

Twenty-four new participants from introductory courses in psychology received course credit for their participation. We discarded the data of two participants who gave the same response on all trials. Stimuli were presented six times each, in blocks organized into two conditions: (a) audio-visual, and (b) audio-alone. The seven blocks included three blocks of display:gesture (72 trials), three blocks of display: text (72 trials), and one block of audio-alone (36 trials) for a total of 180 trials per participant.

**Results and Discussion**

For the sake of brevity, we focus our analysis mostly on assessing whether the text condition affects duration ratings. Figure 16 summarizes the results. It shows that the visual influence is larger with the videos than the text.

**The difference between the displays.** The 4.7 (±2.9) point effect of the text was negligible (95% CI: [−0.9, 10.5], p = 0.1), whereas the 10.5 (±2.9) point effect of the video
(95% cr: [4.7, 16.2], $p = 0.002$) tell us that it affected the ratings. The latter was $5.7 \pm 2.5$ points larger (95% cr: [0.8, 10.5], $p = 0.02$) than the effect of the text. (There is only weak evidence that the ratings that the ratings in the text condition were greater than 0: $p = 5 \times 0.5^6 = 0.08$.)

We conclude that it is improbable that the illusion is affected, let alone accounted for, by bias.

Vatakis and Spence offer three hypotheses to account for their conclusion that the unity assumption does not apply to non-speech stimuli: (a) The “speech is special” hypothesis; (b) overlearning of the relation between articulatory classes and facial movements; (c) that auditory and visual streams for speech stimuli require greater temporal correlation than Vatakis and Spence’s conclusions leave us at loss to account for the Schutz and Lipscomb illusion. However the observation that humans readily detect causal relationships within and across modalities (Fisher, 1962; Guski & Troje, 2003) may hold the solution to our puzzle.

**General Discussion**

We summarize our studies in Table 1 and Figure 17. We used the Schutz–Lipscomb illusion to test the binding by causality hypothesis. In support of this hypothesis we found that: (a) The visible percussive gesture only affects the perceived duration of sounds it could have caused (percussive but not sustained sounds; Experiment 1). (b) The greater the likelihood that the visible gesture caused the sound, the greater the effect (the effect is greater for marimba than piano; Experiment 1). (c) Since the gesture can cause the sound, but not vice versa, the effect is substantial only when the visible impact precedes the sound (Experiment 2). (d) By and large, the only component of the gesture that affects the perceived duration of the sound is the one concurrent with the sound (Experiment 3).

We ruled out the uncertainty hypothesis, which is an application of the optimal integration hypothesis, according to which the the effect of gesture in the Schutz–Lipscomb illusion occurs for sounds whose duration is uncertain. Since percussive sounds decay gradually, and do not have a clear offset, this hypothesis is a plausible account of the illusion. To rule it out, we asked whether the effect of the gesture in the audio-visual conditions in Experiments 1–3 is predicted by the variability of duration judgments — the standard index of perceptual uncertainty — in the audio-alone conditions. It is not.

We also considered a response-bias hypothesis, according to which the effect of gesture in the Schutz–Lipscomb illusion is cognitively mediated rather than perceptual: because — as we showed in Experiment 3 — the long and the short gestures are visibly different, the gesture might trigger the idea of short or long duration and thus affect the rating without affecting perception. Experiment 4 ruled out the most blatant form of such bias, the direct effect of words on duration ratings. Nevertheless it does not address a more subtle form of the hypothesis: that the effect requires relevant gestural input, but that it is not mediated by perception. Indeed in two experiments we found unmistakable evidence of a weak

![Figure 17](image.png)
Experiment 2 and 3, we see that they may not have been caused by response bias, but by binding that requires proximity in time rather than causality. As we pointed out in our review of the relevant literature in our discussion of Experiment 3, there are indeed asymmetries between cross-modal influences when sound precedes visual information and when it follows. But these studies have not ruled out cross-modal effects when the sound comes first.

We also have evidence that one cannot counteract the effects of binding by causality. Our experiments suggest that one does not have the option to voluntarily undo the cross-modal influences that follow from such binding: we gave our participants instructions to ignore the gestures when they rated tone duration. They were able to do so only when it was unlikely that the gesture caused the tone.

Finally, we found that participants were less sensitive to differences in auditory duration in the presence of visual information. In each experiment, the slopes of the linear relation between ratings obtained in the audio-visual condition and the audio-alone condition were less than 1. This reduction in sensitivity may reflect competition for attentional resources between the auditory and visual systems.

To conclude, we must point out some limitations of this work, which suggest directions for future research. (a) We have not shown that other types of causal links can trigger cross-modal binding and hence cross-modal influences. To turn our hypothesis into a theory we would have to show that visual events which seem to cause sustained sounds (the sounds of bowed strings or wind instruments) trigger binding by causality and produce cross-modal influence. Since visual information is known to affect ratings of sung intervals (Thompson & Russo, 2007) and tension in clarinet performances (Vines, Krumhansl, Wanderley, & Levitin, 2006), it is not unlikely that such audio-visual pairs can be found. (b) We do not know whether a visual event that is not seen to cause a percussive sound fails to influence the perception of this sound.

**Final remarks**

Although the origin of the privileged relationship between impact gestures and percussive sounds is not yet clear, we believe that it is a perceptual adaptation. Percussive sounds produced by impacts are common in our environment (branches breaking, rocks falling, footsteps, etc.), and they are important to our perception of it. Our results may also reflect cross-modal Gestalt principles that integrate auditory and visual information into a single “impact event.” Finally, this kind of relationship may be learned from repeated exposure to impact events.

Regardless of its origin, it makes sense for the perceptual system to treat artificial, ecologically unrelated information, such as tone and light pairs, in a default mode in which the best information wins, and to have a privileged override for information that specifies a common cause. This heuristic captures the best of both worlds, defaulting to the stronger modality except when there is a compelling reason to do otherwise. Such a design optimizes information utility rather than information quality.

The discovery of the illusion by Schutz and Lipscomb (2007) was prompted by a debate among two schools of thought about marimba performance. Some thought that by modulating one’s gesture one could lengthen the physical duration of a marimba sound to achieve an aesthetic effect; others claimed that one could not because it was physically impossible. As is so often the case, those who were wrong in theory were right in practice: as it happens, the physically impossible may be aesthetically powerful.

**References**


